Predicción del precio del café Colombiano por medio de sistemas de inferencia difusos

Karen Martinez
Facultad de Ingeniería
Universidad Distrital Francisco José de Caldas
kgmartinezm@correo.udistrital.edu.co

Angie Rodríguez
Facultad de Ingeniería
Universidad Distrital Francisco José de Caldas
anprodriguezh@correo.udistrital.edu.co

Dayana Torres
Facultad de Ingeniería
Universidad Distrital Francisco José de Caldas
dktorresh@correo.udistrital.edu.co

***Abstract*— In this paper we present the design of a fuzzy forecaster for the monthly price of Colombian coffee in a thirteen-year time window. This is an internationally recognized index that provides coffee producers and dealers with useful information on the management of their annual work. This community is interested in forecasting the index to increase its profits. Four calibration methods are used to tune up the fuzzy forecaster: experience-based design, simple genetic algorithms, coevolutionary algorithms and stochastic hill climbing. The results obtained from these methods in a two-year validation time window are compared in terms of several indices that show the best forecaster is achieved by the coevolutionary algorithm.**

se presentan cuatro métodos para el diseño de sistemas de inferencia difusos con el fin de predecir el precio mensual del café colombiano. El primero de ellos se basa en el conocimiento empírico del problema, los dos siguientes emplean algoritmos de optimización, mientras que el segundo método utiliza un algoritmo genético, el tercero utiliza uno coevolutivo, el último método se basa en un algoritmo estocástico hill climbing. Se realiza una comparación de los cuatro métodos evaluando el resultado de la predicción por medio de diferentes estadísticos, obteniendo la mejor solución en estos términos, por último, se caracteriza.

I. INTRODUCCIÓN

Desde la segunda mitad del siglo XIX el café ha estado ligado al desarrollo histórico colombiano, no sólo por haberse constituido durante más de un siglo en el principal producto de exportación nacional, sino también porque a través de él se ha generado una cultura que, hoy por hoy, lo ha convertido en símbolo de nuestra identidad colombiana por todo el mundo.[1]

El sector cafetero ha enfrentado diversos problemas durante las últimas décadas; entre ellos se cuentan los vaivenes de los precios internacionales los cuales tienden más a la alza; y el desplazamiento del grano como primera fuente de divisas del país por otros productos no agrícolas. A la fecha, se han realizado varios estudios encaminados a modelar el comportamiento del precio del café y es necesario determinar si existe alguna ventaja derivada del uso de modelos no lineales para representar la dinámica que siguen los precios.

Por esta razón se busca implementar un sistema de inferencia difuso, el cual se caracteriza por ser no lineal, y de esta forma buscar predecir el precio de café colombiano, brindado información que podría ser aprovechada por especialistas en el tema.

II. DESCRIPCIÓN DEL PROBLEMA

II-A. *Base de datos*

Se tomó la base de datos correspondiente al precio mensual del café Colombiano del Centro de estudios Latinoamericanos (CESLA) de la Universidad Autónoma de Madrid [4], la cual comprende los precios desde enero de 1995 hasta diciembre de 2016 expresados en USD.

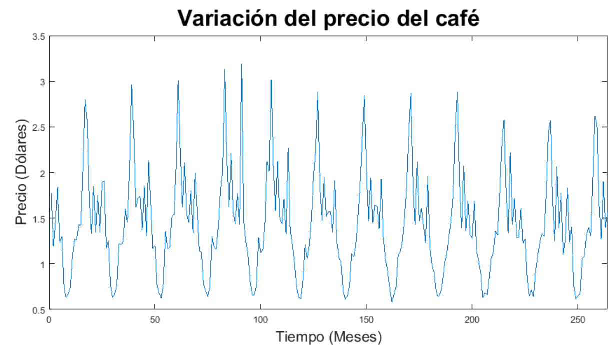


Fig. 1. Serie de tiempo para el precio del café

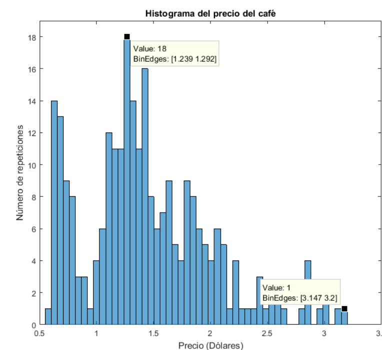


Fig. 2. Histograma para la serie de tiempo

Se posee un total de 264 datos, la gráfica de la serie de tiempo se muestra en la figura 1. Se realiza un histograma

Entrada	α_0
Precio global del café	0.9582
Precio pasado del café	0.9484
Precio del dólar	-0.3890
Producción	-0.2621

TABLE I
COEFICIENTES DE CORRELACIÓN

para identificar el comportamiento de los datos, con el fin de representar su distribución en frecuencia, este se muestra en la figura 2, se observa que el dato que más se repite esta en el rango de 1.239 a 1.292 USD, la media del problema se encuentra en 1.44 y la mediana en 1.36.

Se puede ver que en los primeros y últimos meses del año el precio tiende a caer y mientras que en los meses de la mitad, el precio aumenta, esto puede ser debido a varios factores como la producción y la oferta del café.

El precio del café colombiano podría variar según el comportamiento de algunas variables tales como: el precio global del café, el clima, la producción, el precio del dolar, entre otras. Para el análisis de su comportamiento en este trabajo se utilizaron tres entradas exógenas las cuales corresponden a la producción, el precio global del café y el precio del dólar, debido a que el clima es una entrada linealmente dependiente de la producción puede descartarse para el análisis. Otra entrada a tener en cuenta son los precios pasados, ya que al ser un problema de predicción estos pueden influir en el precio actual.

II-B. *Análisis de las entradas del sistema de inferencia difuso*

El criterio a tener en cuenta para la selección de las entradas es la correlación entre estas y el precio del café. Los coeficientes de correlación se observan en la tabla 1. De esta tabla se puede inferir que las entradas que más influyen en el comportamiento del precio del café son el precio global y el precio pasado.

Se realizó la autocorrelación lineal para la serie de tiempo del precio del café colombiano la cuál se muestra en la figura 3. Se observa que la señal podría presentar memoria, lo cuál hace referencia a una dependencia temporal de los datos.

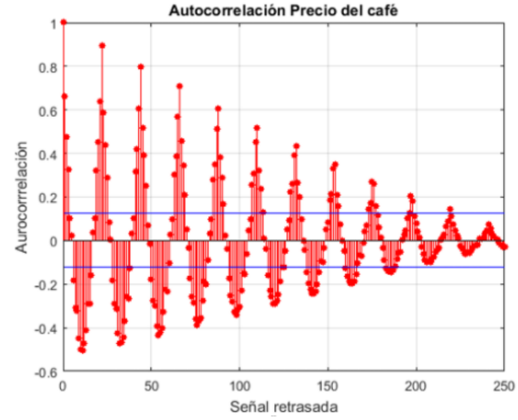


Fig. 3. Autocorrelación de la serie de tiempo del precio del café

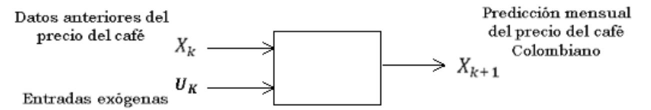


Fig. 4. Modelo de predicción del precio del café Colombiano

En la figura 4 se aprecia el diseño del sistema para la predicción del precio del café colombiano. Este está basado en el modelo ARMAX, el cual tiene en cuenta las entradas exógenas del problema y los datos pasados del mismo.

III. CONFRONTACIÓN DE MÉTODOS

Con el propósito de encontrar un sistema de inferencia difuso que presente una buena aproximación a la predicción del precio del café, se proponen cuatro métodos, el primero de ellos se basa en el conocimiento empírico del problema, los dos siguientes emplean algoritmos de optimización y el último utiliza el algoritmo hill climbing. A continuación se presenta una breve descripción y los resultados para cada uno de ellos.

III-A. *Primer método de diseño del sistema de inferencia difuso: Conocimiento empírico sobre el problema*

Se realizó una revisión bibliográfica con el fin de obtener información acerca de la naturaleza del problema y de las variables seleccionadas para el diseño del sistema de inferencia difuso. De la información obtenida se identificaron las características de cada una de las variables de entrada, tales como el rango de valores y el comportamiento que manifiestan durante el período de tiempo analizado y el conjunto numérico al cual pertenecen (reales, enteros, etc). A partir del anterior análisis se establecen los conjuntos difusos para cada entrada.

Teniendo en cuenta la revisión bibliográfica y la información presentada en la tabla I, se observa que las dos entradas más relevantes son el precio global del café y el precio pasado,

por lo tanto se propone clasificarlas a partir de cinco etiquetas linguísticas las cuales son: Muy alto, alto, medio, bajo y muy bajo, para la creación de estos conjuntos difusos se elige la forma triangular y se distribuye de manera uniforme dentro de los rangos de las variables. Para el precio del dolar y la producción se propuso la clasificación únicamente con dos etiquetas linguísticas: alto y bajo y para su distribución dentro del rango de las variables correspondientes se escogieron los conjuntos Gaussianos.

Con los conjuntos difusos creados se realizó una base de 21 reglas, luego se procedió a implementar el sistema difuso y evaluar su resultado con un conjunto de datos del 10 % de la base de datos total.

Para evaluar el error que presenta se escogió el error cuadrado promedio, este evalua la varianza que tiene la señal obteniada con respecto a la real, y se utiliza más que todo en problema de predicción,ya que este proporciona información relevante para observar si se esta realizando un buen predictor; si el valor de este es cercano a 0, es mejor el predictor.

III-A.1. Resultados: El sistema se evaluó con el 10 % de la base de datos obteniendo el siguiente resultado:

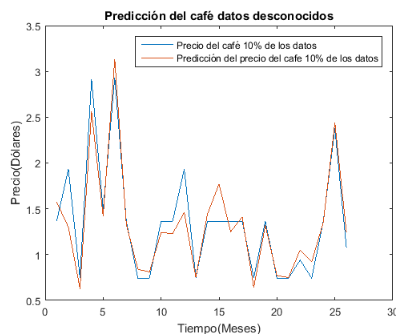


Fig. 5. Resultado obtenido después de evaluar el sistema con el 10 % de la base de datos

Luego de evaluar el resultado se obtienen los siguientes datos : correlación = 93.93 % , RMSE= 2.183 % y MSE=1 .2 4 %.

III-B. Segundo método de diseño del sistema de inferencia difuso: Aprendizaje supervisado

Para este segundo método se aplicó el sistema adaptativo de inferencia neuro difusa (ANFIS), el cual se caracteriza por que el sistema va .ªprendiendoçon la base de datos, mientras un supervisor indica cuantas veces se le debe mostrar la base de datos al sistema para que este pueda estar en la capacidad de aprender. La idea fundamental de los sistemas ANFIS consiste en dividir en dos o más regiones cada una de las variables de entrada o regresores. De esta forma, el dominio del problema resulta dividido en un conjunto de regiones que surge de la intersección de las regiones en que ha sido dividido cada regresor.

Para el modelo ANFIS, se debe dejar que la máquina observe la base de datos cuantas veces sea necesario para que aprenda, pero también se observa que si esta se deja expuesta de sobremanera a la observación de la base de datos, el sistema podría "sobre-aprenderz así especializarce en el conocimiento de dicha base, pero cuando se ponga a prueba con datos desconocidos, esta no estaría en la capacidad de reconocerlos. De esto se puede detallar que exixte una combinación entre la necesidad de mostrarle la base de datos al sistema, y el chance de que esta no sobre-aprenda.

En este caso, se dejaron 1500 experimentos con 100 iteraciones, es decir, que se le mostró la base de datos al sistema 100 veces y de esto se realizaron 1500 experimentos; este proceso se repitió cambiando el numero de reglas, con el fin de obtener el menor número posible, teniendo en cuenta que presentara el menor error. Para conocer el error, se caracterizó el error de entrenamiento del sistema, el cual se hizo con un 90 % de la base de datos, y luego de esto se caracterizó el error de validación el cual se realizó con el 10 % de datos restantes, los cuales eran desconocidos para el sistema y así se podía observar de mejor forma su comportamiento con relación al aprendizaje.En este método se puede observar que aún conociendo la base de datos, los conjuntos difusos y la formulación de las reglas son confusas y no son interpretables.

III-B.1. Resultados: El sistema se evaluó con el 10 % de la base de datos obteniendo el siguiente resultado:

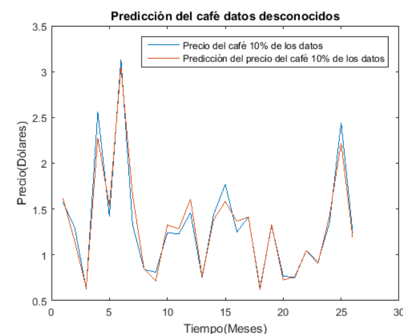


Fig. 6. Resultado obtenido después de evaluar el sistema con el 10 % de la base de datos

Luego de evaluar el resultado se obtienen los siguientes datos : correlación = 97.934 % , RMSE= 0.76 % y MSE=0.15 %

III-C. Tercer método para diseñar el sistema de inferencia difuso: Sistemas evolutivos

En este método se utilizó como ejemplo los sistemas coevolutivos, los cuales tienen en cuenta las poblaciones exteriores que pueden afectar a la población de estudio, en nuestro caso, las entradas exógenas que tiene el problema. Los algoritmos evolutivos se caracterizan porque de estos se seleccionan los mejores "padres"que darán paso a la siguiente generación,posterior a esto se cruzan y se da una pequeña mutación, la cual en unos casos se puede dar como en otros no,obteniendo así la población actual. De esta forma se realiza la semejanza con el problema y la base de datos. Pero a diferencia del algoritmo evolutivo tradicional, en la evaluación de cada individuo para seleccionar cuales darán origen a la

siguiente generación no depende exclusivamente del individuo, sino que este se da en el contexto de la interacción con otros, ya sea de la misma población o de otras. Las posibilidades de supervivencia o reproducción de un individuo pueden variar en función de otros individuos, con lo que se puede decir que la evaluación en un algoritmo coevolutivo es subjetiva. De estos algoritmos se debe hallar una función objetivo con los índices de error más importantes para el problema, de los que se escogieron: El RMSE (Raiz del error cuadrático medio): Este representa la desviación estándar la cual puede ser interpretada como una medida de incertidumbre. La desviación estándar de esas medidas es de vital importancia: si la media de las medidas está demasiado alejada de la predicción (con la distancia medida en desviaciones estándar), entonces se considera que las medidas contradicen la teoría. El MSE (Error cuadrático medio): Se usa el MSE como uno de los índices para la función objetivo debido a que este proporciona información relevante para observar si se esta realizando un buen predictor; si el valor de este es cercano a 0, es mejor el predictor. Esta mide el promedio de los errores al cuadrado, es decir, la diferencia entre el predictor y lo que se predice. De esta se halla el mejor individuo por experimento y se evalua; cabe aclarar que como en el método anterior se usó 90 % de la base de datos para evolución y 10 % para validar. Como en el método anterior se realizaron 150000 pruebas, en este método se busca igualar este mismo número de pruebas con el fin de comparar. Por ello se realizan 100 generaciones de a 30 individuos cada una, y a con estas se realizan 50 experimentos, con la base de 5 reglas. Una vez vez obtenidos los mejores individuos por experimento se caracterizan con el 10 % de la base de datos, los cuales son datos desconocidos, y de allí se halla al mejor individuo, el cual se va a confrontar más adelante con los otros métodos.

III-C.1. Resultados: El sistema se evaluó con el 10 % de la base de datos obteniendo el siguiente resultado:

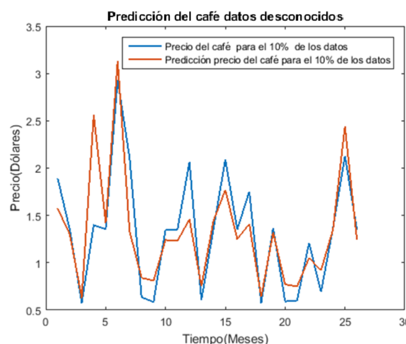


Fig. 7. Resultado obtenido después de evaluar el sistema con el 10 % de la base de datos

Luego de evaluar el resultado se obtienen los siguientes datos : correlación = 83.45 % , RMSE= 0.00021 % y MSE=0.0287 %

III-D. Cuarto método para diseñar el sistema de inferencia difuso: Aletoriedad

Para el método aleatorio, se crearon sistemas difusos al azar y se evaluaron con la base de datos bajo la función objetivo del anterior método, esto con el fin de garantizar que los sistemas difusos se evalúan bajo las mismas condiciones para obtener el mejor de los que se han creado por método. En este método se debe garantizar que los experimentos se recrean bajo las mismas condiciones para poder hacer una caracterización estadística del error,tomando el error por experimento y de allí verificar cual presenta la aproximación más adecuada a la solución del problema. En este método se fijan un número de 5 reglas, 4 entradas 2 de ellas con 4 conjuntos difusos y las otras 2 con 2 conjuntos difusos, como se deben evaluar tantas posibilidades como en los experimentos anteriores se realizan 30 experimentos cada uno con 5000 iteraciones, en cada una de las iteraciones se creó un sistema difuso y se evaluó, los resultados se fueron comparando para ir eligiendo el mejor de cada experimento, una vez obtenidos los mejores sistemas difusos de cada experimento se evalúan con la función objetivo y de nuevo se toma el que menor error presente.

III-D.1. Resultados: El sistema se evaluó con el 10 % de la base de datos obteniendo el siguiente resultado:

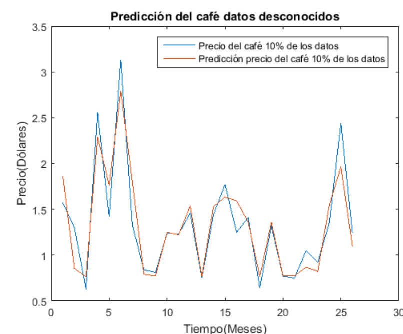


Fig. 8. Resultado obtenido después de evaluar el sistema con el 10 % de la base de datos

Luego de evaluar el resultado se obtienen los siguientes datos : correlación = 92.29 % , RMSE= 0.0288 % y MSE=0.00021 %

IV. CARACTERIZACIÓN DEL MEJOR SISTEMA OBTENIDO

Se elige el sistema difuso del cuarto método como el mejor sistema difuso creado, puesto que presenta los errores de MSE y RMSE más bajos y la correlación más alta con respecto al resultado del primer método, además de contar con el menor número de reglas posible y una sencilla interpretación de los conjuntos difusos, a continuación se presenta la caracterización del sistema y la descripción del mismo.

A continuación, se presentan las gráficas de caracterización del error del cuarto método puesto que este es el que arroja como resultado el mejor sistema de inferencia de los 4 métodos:

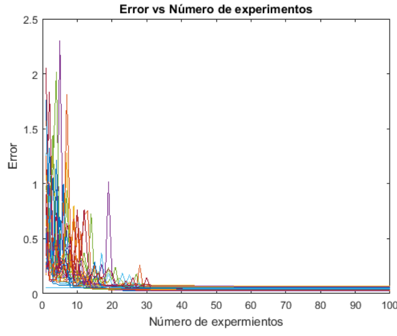


Fig. 9. Error vs número de experimentos

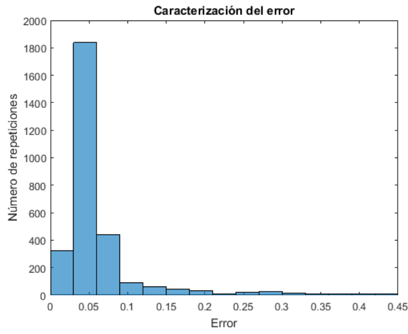


Fig. 10. Histograma del eror

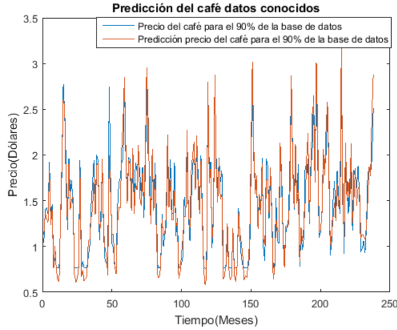


Fig. 11. Resultado obtenido después de evaluar el sistema con el 90 % de la base de datos

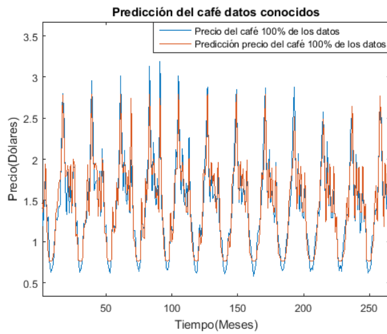


Fig. 12. Resultado obtenido después de evaluar el sistema con el 100 % de la base de datos

Los conjuntos difusos de este sistema son conjuntos que tienen una forma gaussiana y se pueden interpretar mediante etiquetas lingüísticas como: muy bajo, bajo, medio, alto en el caso de las entradas que tienen 4 conjuntos difusos, y alto y bajo para el caso de las entradas que cuentan con 2 conjuntos difusos, se puede decir que esto los hace fáciles de entender.

IV-A. Oscilador caótico

En el estudio del problema se observó que el comportamiento del sistema se asemeja a un oscilador caótico. Uno de los osciladores caóticos más conocidos corresponde al circuito de Chua [5], el cual puede manejarse de una manera simple y aún así, presentar robustez; este circuito puede ser implementado a partir de cinco elementos: dos capacitores, un inductor , una resistencia lineal y una resistencia no lineal más conocida como el diodo de Chua (Figura 5), en este problema se podría cambiar la resistencia no lineal por precio del café.

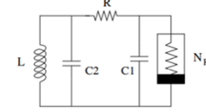


Fig. 13. Circuito para el modelamiento del Oscilador caótico de Chua

Otro de los osciladores caóticos más conocidos es el atractor de Lorenz el cual está descrito por las siguientes ecuaciones, las cuales estan basadas en las ecuaciones dinámicas de la atmósfera terrestre.[6]:

$$\dot{X} = \alpha(Y - X) \tag{1}$$

$$\dot{Y} = (b - Z)X - Y \tag{2}$$

$$\dot{Z} = XY - cZ \tag{3}$$

En donde a b y c son constantes, y XY y XZ corresponden a las no linealidades responsables de generar el caos.

Para la generación del diagrama de fase del oscilador caótico es necesario hacer uso del teorema de Takens, el cual realiza la incrustación de un retraso proporcional a las condiciones bajo las cuales se puede reconstruir un sistema dinámico caótico a partir de una secuencia de observaciones del estado de un sistema dinámico.

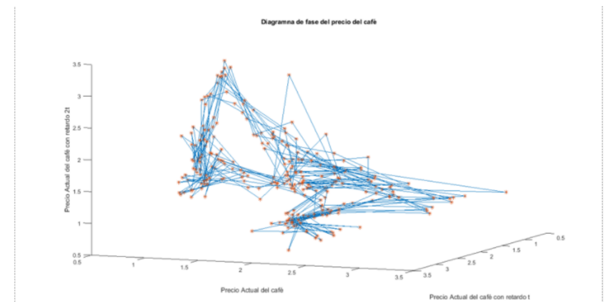


Fig. 14. Diagrama de fase para el precio del café

El comportamiento observado se asemeja a un atractor de Lorenz, sin embargo, el atractor obtenido presenta cambios abruptos, puesto que los cambios en los precios son fuertes.

V. CONCLUSIONES

- Se puede observar en el primer método que aunque la reglas formuladas y los conjuntos difusos son totalmente entendibles, el número de reglas es bastante alto en comparación con las que se obtuvieron en los otros métodos.
- El segundo método presenta muy buenos resultados y se logra reducir de manera significante el número de reglas, pero la interpretación de los conjuntos difusos y las reglas del sistema difuso no es posible.
- En el caso del tercer y cuarto método los resultados son bastante parecidos y se logra una reducción del número de reglas, además de que en estos casos tanto los conjuntos como las reglas son interpretables.
- El sistema de inferencia difuso elegido como el mejor, presenta una aproximación favorecedora a la solución del problema, esto se evidencia en los resultados de error y en las respuestas que se dan al evaluarlo con la base de datos.

REFERENCES

[1] VELASQUEZ HENAO, Juan David and ALDANA DUMAR, Mario Alberto. *MODELLING OF THE COLOMBIAN COFFEE PRICE IN THE NEW YORK STOCK EXCHANGE USING ARTIFICIAL NEURAL NETWORKS.* Rev.Fac.Nal.Agr.Medellín. (2007, vol.60, n.2, pp.4129-4144. ISSN 0304-2847.)

[2] PEREZ RAMIREZ, Fredy Ocaris *Modelación de la volatilidad y pronóstico del precio del café.* Rev. ing. univ. Medellin. (2006, vol.5, n.9, pp.45-58. ISSN 1692-3324.)

[3] GARCÍA MARTÍN, Ismael *ANÁLISIS Y PREDICCIÓN DE LA SERIE DE TIEMPO DEL PRECIO EXTERNO DEL CAFÉ COLOMBIANO UTILIZANDO REDES NEURONALES ARTIFICIALES.* Universitas Scientiarum, [S.l.] (v. 8, p. 45-50, jul. 2003. ISSN 2027-1352.)

[4] CESLA - CENTRO de ESTUDIOS LATINOAMERICANOS.Facultad CC. EE. y EE. (Mod. E-XIV) Universidad Autónoma de Madrid. *Base de datos del precio promedio del café. (USD$/Libra)* CP 28049 - Cantoblanco, Madrid - España (Correspondiente a los años desde 1995 al 2017)

[5] M. A. Duarte-Villaseñor, E. Tlelo-Cuautle y J. M. García-Ortega *Modelado y Simulación de un Oscilador Caótico usando MatLab*. IEEE LATIN AMERICA TRANSACTIONS, VOL. 5, NO. 2, MAY 2007 95

[6] A.S. Elwakil , S. Ozoguz and M.P. Kennedy *Creation of a complex butterfly attractor using a novel Lorenz-Type system*. IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications (Volume: 49, Issue: 4, Apr 2002)